# Targeted Next-generation Sequencing and Bioinformatics Pipeline to Evaluate Genetic Determinants of Constitutional Disease

Video Article

# Targeted Next-generation Sequencing and Bioinformatics Pipeline to Evaluate Genetic Determinants of Constitutional Disease

Allison A. Dilliott[1,2], Sali M.K. Farhan[3], Mahdi Ghani[4], Christine Sato[4], Eric Liang[5], Ming Zhang[4], Adam D. McIntyre[1], Henian Cao[1], Lemuel Racacho[6,7], John F. Robinson[1], Michael J. Strong[1,8], Mario Masellis[9,10], Dennis E. Bulman[6,7], Ekaterina Rogaeva[4], Anthony Lang[10,11], Carmela Tartaglia[4,10], Elizabeth Finger[12,13], Lorne Zinman[9], John Turnbull[14], Morris Freedman[10,15], Rick Swartz[9], Sandra E. Black[9,16], Robert A. Hegele[1,2]

[1]Robarts Research Institute, Schulich School of Medicine and Dentistry, Western University

[2]Department of Biochemistry, Schulich School of Medicine and Dentistry, Western University

[3]Analytic and Translational Genetics Unit, Center for Genomic Medicine, Harvard Medical School, Massachusetts General Hospital, Stanley Centre for Psychiatric Research, Broad Institute of MIT and Harvard

[4]Tanz Centre for Research in Neurodegenerative Diseases, University of Toronto

[5]School of Medicine, Faculty of Health Sciences, Queen's University

[6]Faculty of Medicine, Department of Biochemistry, Microbiology and Immunology, University of Ottawa

[7]CHEO Research Institute, Faculty of Medicine, University of Ottawa

[8]Department of Clinical Neurological Sciences, Western University

[9]Division of Neurology, Department of Medicine, Sunnybrook Health Sciences Centre, University of Toronto

[10]Division of Neurology, Department of Medicine, University of Toronto

[11]Morton and Gloria Shulman Movement Disorders Centre, Toronto Western Hospital

[12]Department of Clinical Neurological Sciences, Schulich School of Medicine and Dentistry, Western University

[13]Parkwood Institute, St. Joseph's Health Care

[14]Department of Medicine, Division of Neurology, McMaster University

[15]Division of Neurology, Department of Medicine, Baycrest Health Sciences

[16]Canadian Partnership for Stroke Recovery Sunnybrook Site, Sunnybrook Health Science Centre, University of Toronto

Correspondence to: Robert A. Hegele at hegele@robarts.ca

URL: https://www.jove.com/video/57266
DOI: doi:10.3791/57266

Keywords: Genetics, Issue 134, Next-generation sequencing, targeted sequencing, resequencing, variant calling, variant annotation, constitutional disease

Date Published: 4/4/2018

Citation: Dilliott, A.A., Farhan, S.M., Ghani, M., Sato, C., Liang, E., Zhang, M., McIntyre, A.D., Cao, H., Racacho, L., Robinson, J.F., Strong, M.J., Masellis, M., Bulman, D.E., Rogaeva, E., Lang, A., Tartaglia, C., Finger, E., Zinman, L., Turnbull, J., Freedman, M., Swartz, R., Black, S.E., Hegele, R.A. Targeted Next-generation Sequencing and Bioinformatics Pipeline to Evaluate Genetic Determinants of Constitutional Disease. *J. Vis. Exp.* (134), e57266, doi:10.3791/57266 (2018).

## Abstract

Next-generation sequencing (NGS) is quickly revolutionizing how research into the genetic determinants of constitutional disease is performed. The technique is highly efficient with millions of sequencing reads being produced in a short time span and at relatively low cost. Specifically, targeted NGS is able to focus investigations to genomic regions of particular interest based on the disease of study. Not only does this further reduce costs and increase the speed of the process, but it lessens the computational burden that often accompanies NGS. Although targeted NGS is restricted to certain regions of the genome, preventing identification of potential novel loci of interest, it can be an excellent technique when faced with a phenotypically and genetically heterogeneous disease, for which there are previously known genetic associations. Because of the complex nature of the sequencing technique, it is important to closely adhere to protocols and methodologies in order to achieve sequencing reads of high coverage and quality. Further, once sequencing reads are obtained, a sophisticated bioinformatics workflow is utilized to accurately map reads to a reference genome, to call variants, and to ensure the variants pass quality metrics. Variants must also be annotated and curated based on their clinical significance, which can be standardized by applying the American College of Medical Genetics and Genomics Pathogenicity Guidelines. The methods presented herein will display the steps involved in generating and analyzing NGS data from a targeted sequencing panel, using the ONDRISeq neurodegenerative disease panel as a model, to identify variants that may be of clinical significance.

## Video Link

The video component of this article can be found at https://www.jove.com/video/57266/

## Introduction

As defining the genetic determinants of various conditions takes on a higher priority in research and in the clinic, next-generation sequencing (NGS) is proving to be a high-throughput and cost-effective tool to achieve these goals[1,2,3]. For almost 40 years, Sanger sequencing had been the gold standard for identifying genetic variants[4]; however, for diseases with genetic heterogeneity or unknown genetic etiology, many possible candidate genes must be evaluated, often concurrently. In this context, Sanger sequencing becomes expensive and time-consuming. However, NGS involves massive parallel sequencing of millions of DNA fragments, allowing for a cost and time efficient technique to simultaneously detect a wide range of genetic variation across various regions of the genome.

There are three types of NGS for sequencing DNA: 1) whole-genome sequencing (WGS), 2) whole-exome sequencing (WES), and 3) targeted sequencing[5]. WGS evaluates the entire genomic content of an individual, while WES involves sequencing only the protein-coding regions of the genome[6]. Targeted sequencing, in contrast, focuses on specific regions of the genome based on relatively few specific genes linked by common pathological mechanisms or known clinical phenotype. Either the exons or introns, or any intergenic regions of a gene or specific group of genes can be specified using this approach. Therefore, targeted sequencing can be an excellent approach when there is already a foundation of candidate genes known to be associated with the disease of interest. Targeting specific regions of the genome allows for elimination of superfluous and irrelevant genetic variation that can cloud or distract from clinical interpretation. While WGS and WES both produce a large amount of high-quality data, the amount of data can be overwhelming. Not only does this large amount of data require computationally intensive bioinformatics analysis, but data storage can frequently present problems[7]. This challenge of data storage also adds additional costs to both WGS and WES, which is often not initially considered when calculating the expense of sequencing. Further, although it is decreasing, the cost of WGS and WES remain relatively high. Targeted sequencing can be a more cost-efficient option, particularly when sequencing of a large number of individuals is required.

The Ontario Neurodegenerative Disease Research Initiative (ONDRI) is a multi-platform, provincial-wide, observational cohort study characterizing five neurodegenerative diseases, including: 1) Alzheimer's disease and mild cognitive impairment, 2) amyotrophic lateral sclerosis, 3) frontotemporal dementia, 4) Parkinson's disease, and 5) vascular cognitive impairment[8]. The ONDRI genomics subgroup is aiming to elucidate as part of the baseline characterization of this cohort the often discounted, yet extremely important genetic landscape of these phenotypically and genetically heterogeneous diseases. Neurodegenerative diseases are thus appropriate candidates for NGS methodologies and for targeted sequencing in particular.

We have custom-designed a targeted NGS panel, ONDRISeq, to sequence 528 participants involved in ONDRI for the protein-coding regions of 80 genes that have been previously associated with the five diseases of interest. With this methodology, we are able to harness the high-quality NGS data in a focused and efficient manner. The design and validation of the ONDRISeq panel with multiple concordance studies has been previously described, for which the ONDRISeq panel was able to identify novel, rare variants of possible clinical significance in 72.2% of 216 cases used for panel validation[9]. Although NGS technology has advanced rapidly and remarkably in recent years, many researchers face a challenge when processing the raw data into a list of usable, annotated variants[10]. Further, interpretation of the variants can be complex, especially when faced with many that are rare or novel[11].

Here, we describe in a step-by-step manner, the methodology of targeted NGS and the associated bioinformatics workflow required for resequencing, variant calling, and variant annotation using the ONDRISeq study as an example. After the generation of NGS data, raw sequencing files must be aligned to the human reference genome in order to accurately call variants. Variants must then be annotated in order to perform subsequent variant curation. We will also explain our implementation of the American College of Medical Genetics' Standards and Guidelines to accurately classify variant pathogenicity.

## Protocol

For the purposes of ONDRI, ethics protocols and informed consent were obtained based on the Research Ethic Boards at Baycrest Centre for Geriatric Care (Toronto, Ontario, Canada); Centre for Addiction and Mental Health (Toronto, Ontario, Canada); Elizabeth Bruyère Hospital (Ottawa, Ontario, Canada); Hamilton General Hospital (Hamilton, Ontario, Canada); London Health Sciences Centre (London, Ontario, Canada); McMaster (Hamilton, Ontario, Canada); The Ottawa Hospital (Ottawa, Ontario, Canada); Parkwood Hospital (London, Ontario, Canada); St Michael's Hospital (Toronto, Ontario, Canada); Sunnybrook Health Sciences Centre (Toronto, Ontario, Canada); and University Health Network-Toronto Western Hospital (Toronto, Ontario, Canada).

# 1. DNA Isolation from Human Blood Samples

1. **Collect samples from sequencing participants in accordance with appropriate ethics protocols and informed consent.**
    1. To obtain DNA of high quality, draw blood samples for the purposes of extraction.
    NOTE: DNA can also be extracted from saliva or buccal cells, ensuring that an appropriate DNA extraction kit is used.
    2. If extracting from blood, to obtain a high yield of DNA, collect the sample in three 4 mL EDTA K2 tubes, providing a sample of total volume ~12 ml.
    3. Centrifuge blood samples for 20 min at 750 x g to fraction into an upper phase of plasma, thin, middle phase of leukocytes, and a bottom phase of erythrocytes.
2. Remove the plasma from the blood sample by pipetting it off the top of the sample with a disposable transfer pipette. Appropriately discard the plasma or dispense into multiple 500 µL aliquots for storage at -80 °C for future biochemical analyses. Ensure that a new, sterile pipette is used for each sample.
3. Extract DNA from the blood sample with a blood extraction kit[12] (**Table of Materials**) according to manufacturer's instructions.
NOTE: If a sample of the volume described above is obtained, ~3 mL of leukocytes will be obtained to use in the DNA extraction.

4. Measure initial DNA concentration in ng/µL using a full-spectrum spectrophotometer[13] (**Table of Materials**), according to manufacturer's instructions.
5. Proceed directly to step 2. Alternatively, store DNA at 4 °C.

## 2. Sequencing Library Preparation

1. **Perform serial dilutions on the DNA samples over the course of three days to obtain a final concentration of 5.0 ± 1.0 ng/µL.**
    1. Dilute 1 M Tris buffer pH 8.5 to 10 µM with deionized water.
    NOTE: The volume diluted will depend on the number of DNA samples that will need to be diluted in the subsequent steps.
    2. If performing the DNA dilution directly after step 1.4, proceed to the following step. If not on the same day, measure the DNA concentration as was done in step 1.4.
    3. Based on the concentration measured, dilute 40 µL of the DNA to ~10 ng/µL using 10 µM Tris buffer pH 8.5 and allow the sample to sit overnight at 4 °C.
    4. Measure DNA concentration with a fluorometer[14] appropriate for the quantification of DNA (**Table of Materials**), according to manufacturer's instructions.
    NOTE: The concentration of the sample should be >10 ng/µL because of the lower sensitivity of the spectrophotometer used previously.
    5. Based on the concentration measured, dilute 20 µL of the DNA to 10 ng/µL using 10 µM Tris buffer pH 8.5 and allow the sample to sit overnight at 4 °C.
    6. Measure DNA concentration with the fluorometer[14], according to manufacturer's instructions.
    7. Based on the concentration measured, dilute 10 µL of the DNA to 5 ng/µL using 10 µM Tris-HCl pH 8.5 and allow the sample to sit overnight at 4 °C.

2. **Prepare sequencing library according to manufacturer's instructions with the targeted NGS panel's appropriate target enrichment kit[15] (Table of Materials). Ensure that the enrichment kit is appropriate for the NGS platform being used.**
    1. Follow manufacturer's instructions[16] regarding the plexity and pooling of libraries.
    NOTE: For ONDRISeq, libraries are composed of 12 DNA samples, pooled in sets of two, and run on the NGS desktop instrument (**Table of Materials**). The number of samples that can be run in a single reaction will depend on the sequencing kit and platform used.
    2. To achieve higher quality sequencing data, perform the optional step to validate the DNA library quality following tagmentation, described in manufacturer's instruction of the target enrichment kit[15].
        1. Analyze each library in triplicate to ensure the quality of the library yield.
    3. If pooling libraries, measure DNA concentration with the fluorometer[14], according to manufacturer's instructions. Use this concentration to determine the volume of each DNA library to pool to obtain the equimolar ratios recommended by the target enrichment kit being used.

## 3. Next-generation Sequencing

1. **Sequence the library according to the NGS desktop instrument's reagent kit manufacturer's instructions[17,18] (Table of Materials).**
    1. Prepare a sample sheet according to manufacturer's instructions[18] using the appropriate NGS technology software (**Table of Materials**), which will be imported into the NGS desktop instrument's workflow.
    NOTE: For the purposes of ONDRISeq, the application option chosen is 'other', with only the FASTQ files requested (**Figure 1**). Subsequent steps will process these FASTQ files, to allow for full customization of alignment and quality parameters. However, if targeted sequencing is chosen, some NGS instruments are able to process the sequencing data into VCF files themselves. The manufacturer's instructions[18] may be consulted for a full selection of options.
    2. If using a cloud-based computing environment[19] (**Table of Materials**), log in when setting up the sequencing run. Do this after clicking "Sequencing" on the NGS desktop instrument home page.
    3. Following library denaturation[18] according to the manufacturer's instructions, measure DNA library concentration with the fluorometer[14].
    4. Validate the DNA library quality using an appropriate automated electrophoresis system and DNA quality analysis kit[20] (**Table of Materials**), as per manufacturer's instructions.
    5. To convert the DNA concentration from ng/µL to nM, use the following formula[16]
    $$\frac{(concentration\ in\ ng/\mu L)}{(660\ g/mol * average\ library\ size)} * 10^6 = concentration\ in\ nM$$
    NOTE: Average library size will be specific to target enrichment kit being used, and can be obtained from the electrophoresis trace observed in step 3.1.4.
    6. Dilute the sequencing library to a final concentration of 6–20 pM, as appropriate, and volume of 600 µL, according to manufacturer's instructions[21].
    NOTE: The exact concentration needed is dependent on the sequencing kit used. Consult the enrichment kit manufacturer to determine the proper loading concentration.
    7. Dilute, denature, and include a positive control sequencing library[21], according to the manufacturer's instructions.
    8. Keep a log of every sequencing run, which includes the DNA library concentration loaded (pM), the percentage of positive control added, reagent cartridge barcode, application chosen in step 3.1.1, number of index reads, enrichment kit used, read length(s), and the sample sheet name.
    NOTE: The run time of the NGS desktop instrument will depend on the instrument, enrichment kit, and read lengths chosen (4–56 h for the sequencer used in this experiment[22]).

2. Upon completion of the sequencing run, access the "Run Folder", which includes all outputs, by navigating to the NGS desktop instrument home page and clicking "Manage Files". Move the files to a local drive for later access. For a separate option, on a computer, find the files within the cloud-based computing environment[19] by selecting "Runs" on the navigation panel. Select the appropriate sequencing run to navigate to the Run Summary page. Select "Download" to obtain data from the cloud. From the dialog box that appears, select the FASTQ files as the file type to download and click "Download".
3. **From the Run Summary page of the cloud-based computing environment[19,23], navigate to "Charts" to analyze the quality of the sequencing run with the various figures produced by the computing environment. Refer to the manufacturer's instructions[23] for details regarding each figure produced.**
    1. From the Run Charts page, find the figure labeled "Data by Cycle". Under chart, select "Intensity" and under channel select "All Channels". Ensure that this signal intensity plot produced is similar to that produced by sequencing runs performed in the past with the same enrichment kit and NGS desktop instrument.
    NOTE: This reflects the percentage of intensity shown by each base across all 150 cycles. The figure can vary widely depending on the enrichment kit used, which is why it must be compared to past sequencing runs of the same panel.
    2. Select the "Indexing QC" tab within the run navigation panel to find the indexing quality control (QC) histogram, which is on the right-hand side of the page. Ensure that a relatively uniform distribution of % Reads Identified (PF) is observed across all samples.
    NOTE: If any samples have a much lower % Reads Identified (PF) than the rest of the samples, note that the quality of the sequencing data may be affected.

4. **From the Run Summary page of the cloud-based computing environment, navigate to the quality metrics by clicking "Metrics" within the run navigation panel.**
NOTE: Metrics cut-offs will depend on the sequencing platform and enrichment kit being used. There are many metrics that can be utilized based on manufacturer's instructions[23], with the following steps highlighting three that are highly recommended for quality control.
    1. Under "DENSITY (K/MM$^2$)" ensure the cluster density is within the range recommended by the enrichment kit being used (in this case 1,200–1,400 K/mm$^2$).
    2. Under the total "%≥Q30" ensure that the value is ≥85%, reflecting the quality of the sequencing reads.
    NOTE: If lower than this threshold of 85%, note that the quality of the sequencing may be compromised.
    3. Under "ALIGNED (%)" ensure that the value is similar to the % of positive control that was included in the sequencing run.
    NOTE: This acts as a measure of positive control, such that only this percentage of total reads were found to align to the positive control genome. If 1% positive control was used it would be expected that the Aligned (%) would be ~1–5%.

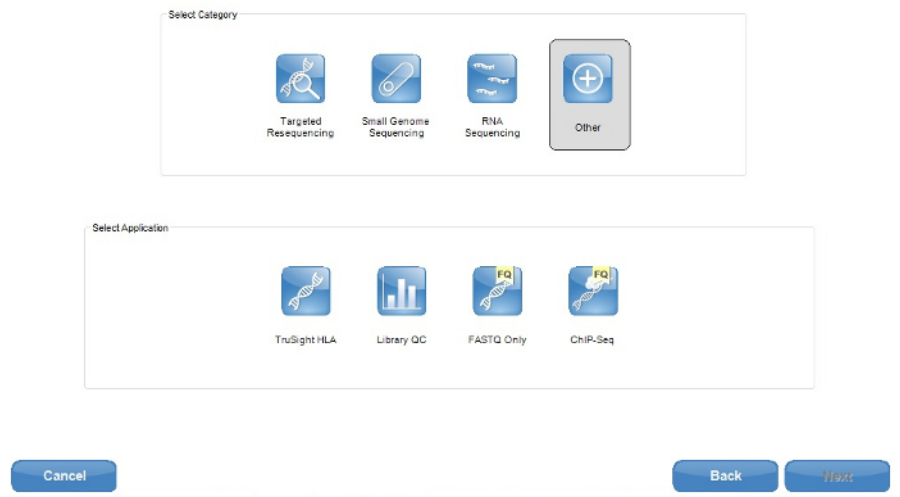



**Figure 1: Screenshot of the NGS technology software's (Table of Materials) sample sheet creator application options.** For the purposes of ONDRISeq, the FASTQ only application is used. However, if the user would like other files produced, such as VCF files, it is recommended that an application within the targeted resequencing category is used. Please click here to view a larger version of this figure.

## 4. Resequencing and Variant Calling

1. For data pre-processing, select appropriate software to align the raw FASTQ files to the human reference genome and to call variants (**Table of Materials**).
2. **Import FASTQ sequencing reads into the data pre-processing software.**
NOTE: For the purposes of ONDRISeq, the 48 FASTQ files produced from a single sequencing run of 24 samples are imported and processed through the software. The number of samples processed at once can vary depending on the needs of the researcher and size of the NGS panel.
    1. Within the "Navigation Area", right click and select "New Folder". Name the folder such that there is clarity as to the sequencing run that was performed.
    2. From the toolbar at the top, select "Import". From the dropdown list of sequencing platforms shown chose the platform with which the sequencing was performed.

NOTE: For the purposes of ONDRISeq, "Illumina" is chosen. However, if using a different sequencing platform consult the manufacturer's instructions for the remainder of the FASTQ importing steps[24].

3. In the dialog box, navigate to and select the FASTQ files from the sequencing run that is being processed. Ensure that the files being imported are stored in and imported from the local drive, if using a computer with multiple servers.
4. From the "General options" of the dialog box, click the box beside "Paired reads" if sequencing used paired end chemistries.
   NOTE: In this case, there should also be two FASTQ samples imported for each sample - one forward and one reverse.
5. From the Paired read information of the dialog box, select "Paired-end (forward-reverse)" if the forward read FASTQ file appears before the reverse read in the file list. If the files appear in the opposite order, select "Mate-pair (reverse-forward)". Set the paired read minimum distance to 1 and maximum distance to 1000, to allow for the detection of small scale structural rearrangements within the sample sequences.
6. From the "Illumina options" of the dialog box, select "Remove failed reads", to remove the reads that failed sequencing. If the NGS desktop instrument de-multiplexed the data before exporting the FASTQ files do not select the "MiSeq de-multiplexing" box.
7. From the "Quality score" dropdown list, select the NGS Pipeline that was utilized for sequencing. Select "Next" at the bottom of the dialog box.
   NOTE: The pipeline used will affect the format of the FASTQ file quality scores. For more information about which pipeline to select, consult the manufacturer's instructions[24].
8. From the new dialog box, select “Save” and “Create subfolders per bath unit to put each sample's FASTQ files into their own individual folder. Select "Next" at the bottom of the dialog box.
9. From the new dialog box, choose the folder that was created in step 4.2.1. This is where the FASTQ files will be imported. Select "Finish" at the bottom of the dialog box and wait until the FASTQ files are imported. Click the "Processes" tab to see the status of the file import.

3. **Design a workflow within the software to perform resequencing and variant calling, according to manufacturer's instructions.**
   NOTE: This workflow can vary based on the needs of the researcher, but the following steps encompass what is included for the purposes of ONDRISeq (**Figure 2**). The steps in this workflow can be applied to other NGS resequencing and variant calling software as appropriate. All bioinformatics processing for the purposes of ONDRI is performed in reference to human reference genome GRCH37/hg19, for consistency of data processing and analysis.
   1. Map the sequencing reads to the reference genome.
      1. When configuring, choose the reference genome as appropriate, ensuring that it is the same reference genome that is used for all bioinformatics steps.
      2. From the masking mode drop-down list select "No masking" so that no regions of the reference sequence are masked.
      3. Use the default mapping options assigned by the software. Review the manufacturer's instructions[24] to verify that this is acceptable based on the purposes of the research.

   2. Include in the workflow local realignment to the human reference genome to resolve any read mapping errors, particularly surrounding insertion-deletion variants.
      1. Use the default local realignment options assigned by the software. Review the manufacturer's instructions[24] to verify that this is acceptable based on the purposes of the research.

   3. Remove duplicated mapped reads produced by PCR within the NGS protocol to reduce the effect of the PCR amplification bias, which may produce false positives[25].
      1. Set the "Maximum representation of minority sequence (%)", based on the needs of the research.
         NOTE: A lenient setting, as used for the purposes of ONDRISeq, is 5%; however, the software's default setting is more stringent 20%. When two reads are very similar, this setting determines if the sequence with fewer read counts should be considered a sequencing error from the PCR amplification bias. Therefore, by setting 5%, the minority read count must be ≤ 5% of the majority read count to be corrected to be identical to the majority read.

   4. Export statistics for the target regions in the form of a coverage summary text file from the read tracks generated in step 4.3.3. Ignore non-specific matches and broken pairs in the settings. Choose a destination on the local drive for these files.
   5. Export a binary sequence alignment map (BAM) file for each sample from the read tracks generated in step 4.3.3. This contains sequence alignment data, if needed in future analyses. Choose a destination on the local drive for these files.
   6. Choose a method of variant detection to call variants within the sequence.
      NOTE: When assumptions can be made about the ploidy of the samples, it is recommended that a fixed ploidy variant detection algorithm be used, as is used for the purposes of ONDRISeq. If this assumption cannot be made, refer to the manufacturer's instructions[24] to determine the best algorithm for the purposes of the research.
      1. When configuring, from the fixed ploidy variant parameters options set the ploidy as appropriate for the sample organism. Set the "required variant probability", or the probability that a variant has been correctly called in order for it to be retained, at 90.0%.
      2. Use the following recommended settings for the general filters: "Minimum coverage" of 10x, "Minimum count" of 2, "Minimum read frequency" of 20%, "Ignore broken pairs", ignore nonspecific matches based on "Reads", and "Minimum read length" of 20.
         NOTE: These parameters are based on the purposes of ONDRISeq. Refer to the manufacturer's instructions[24] to ensure they are appropriate for the research being done.
      3. Use the following recommended settings for the noise filters: "Base quality filters" with a "Neighbourhood radius" mapping quality score of 5, "Minimum central quality" mapping score of 20, and "Minimum neighbourhood quality" mapping score of 15; a "Read direction filter" of 5.0%; and "Relative read direction filter" of 1.0% significance.
         NOTE: These parameters are based on the purposes of ONDRISeq. Refer to the manufacturer's instructions[24] to ensure they are appropriate for the research being done.

   7. Filter the variants that have been called based on their overlap with the targeted panel's target regions as specified by the Browser Extensible Data (BED) file, allowing only variants occurring within the genomic regions selected for the targeted NGS panel to be retained.

NOTE: The BED file will be unique to the targeted NGS panel that is being utilized, based on the regions of the genome that the panel is able to cover.

8. Export a variant report in a variant calling format (VCF) file from the variant track produced in step 4.3.7. Choose a destination on the local drive for these files.
9. Save and install the workflow according to manufacturer's instructions[24], to make it available in the software's "Toolbox". Ensure the workflow is named such that it is clear in the future what NGS panel it is appropriate for.
    1. In the dialog box with the "Exporting reference data" options during installation, set all options to "Bundle".
    2. In the dialog box with the "Install location" options during installation, click "Install the workflow on your local computer".

4. **Run imported FASTQ sequencing read files through the customized bioinformatics workflow designed in step 4.3, according to manufacturer's instructions[24].**
    1. Identify the workflow designed in step 4.3 in the software's "Toolbox" and double-click it.
    2. Within the dialog box that appears, locate the folders of FASTQ files that were imported in step 4.2 within the "Navigation Area". Highlight all folders by selecting them within the "Navigation Area" and then click the box beside "Batch". Use the right-facing arrow to move the files to "Selected elements". Click "Next" at the bottom of the dialog box.
    3. Within the dialog box, review the "Batch overview" to ensure the correct FASTQ files were selected and then click "Next".
    4. Review the following steps of the workflow within the dialog box to ensure the correct files and export locations were selected when designing the workflow in step 4.3: "Map Reads to Reference"; Remove Duplicate Mapped Reads"; "Create Statistics for Target Regions"; "Export BAM"; "Export Tab delimited text"; "Filter Based on Overlap"; and "Export VCF"
    5. Within the final step in the dialog box -"Result handling"- select the option "Save in input folder". Click "Finish" at the bottom of the dialog box.
    NOTE: This means that the files produced for each sample will be placed into the same folder that stores the FASTQ file within the data pre-processing software.

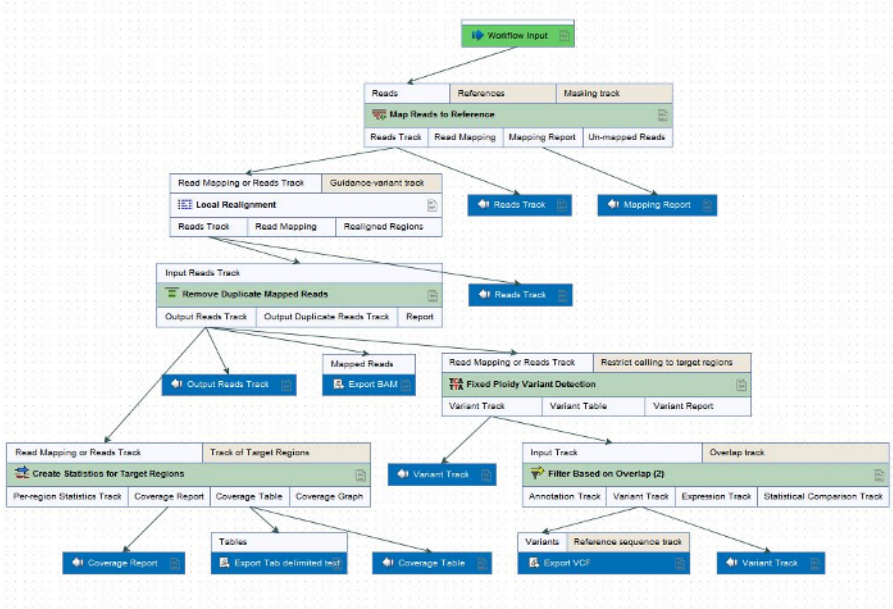



**Figure 2: Workflow for the resequencing and variant calling of FASTQ files within the data pre-processing software (Table of Materials) customized for the purposes of ONDRISeq.** The steps in the workflow can be applied to other NGS resequencing and variant calling software based on the needs of the researcher. Please click here to view a larger version of this figure.

## 5. Variant Annotation

1. **Download and customize the Annotate Variation (ANNOVAR)[26] script to perform variant annotation upon the VCF file of each sample.**
    1. Download the following databases from ANNOVAR to be included as annotations: 1) RefSeq[27] (August 2015 update); 2) dbSNP138[28] (September 2014 update); 3) the Exome Aggregation Consortium[29] (ExAC, version 0.3 November 2015 update); 4) the National Heart, Lung, and Blood Institute Exome Sequencing Project European Cohort[30] (ESP, March 2015 update); 5) the 1000 Genomes Project European Cohort[31] (1KGP, August 2015 update); 6) ClinVar[32] (March 2016 update); and 7) Combined Annotation Dependent Depletion[33] (CADD), Sorting Intolerant from Tolerant[34] (SIFT), and PolyPhen-2[35].
    NOTE: Genome coordinates and all databases referenced by ANNOVAR referred to human genome build GRCh37/hg19. Additionally, the database versions listed are those used for the purposes of ONDRISeq, when downloading the databases use the most up to date versions available.
    2. If desired, customize ANNOVAR to output the complete list of annotated variants, as well as a reduced compilation of annotated variants using the --filter operation[26].

NOTE: The reduced list can be customized based on the needs of the researcher. For the purposes of ONDRISeq, the reduced list of annotated variants does not include variants that occur further than 15 bases from the nearest exon or any variants with a minor allele frequency (MAF) >3% in any of the three databases: 1) ExAC; 2) ESP; and 3) 1KGP. This step is highly recommended.

3. If desired, customize ANNOVAR to single out specific allele calls based on the needs of the researcher[26].
   NOTE: For the purposes of ONDRISeq, ANNOVAR assesses the sequencing calls made for the *APOE* risk alleles rs429358(C>T):p.C130R and rs7412(C>T):p.R176C in order to output the overall *APOE* genotype, of which there are six possible combinations, including: 1) E2/E2; 2) E3/E2; 3) E4/E2; 4) E3/E3; 5) E4/E3; 6) E4/E4. Of these six possible *APOE* genotypes, E4/E4 is the most commonly accepted genetic risk factor for developing late-onset Alzheimer's disease[36].

2. **Query disease mutation databases (Table of Materials) to determine if variants have been previously associated with disease, with reasonable evidence. Consider any variants that have not been previously reported as a novel variant.**
   1. Assess the ANNOVAR annotations from ClinVar, such that the disease-associated variants include any classified as likely pathogenic or pathogenic.

3. Process splicing variants through the *in silico* prediction tools Splicing-based Analysis of Variants[37] (SPANR) and Human Splicing Finder[38] (HSF, version 3.0).
4. If processing a large number of samples, compare the variant calls within each sample to determine which variants are shared by various samples. Do this manually or with a custom-designed script, allowing for the detection of possible sequencing artifacts and contamination events.
   NOTE: For the purposes of ONDRI, a custom script is used to annotate the ANNOVAR output files by comparing them to one another. The script incorporates an annotation, per variant, with the subject ID of any other samples harboring the same variant, otherwise termed the variant's history in the study cohort.
5. Classify variants based on the American College of Medical Genetics (ACMG) Pathogenicity Guidelines[39], assigning each variant a classification as one of the following: 1) pathogenic; 2) likely pathogenic; 3) variant of uncertain significance; 4) likely benign; or 5) benign.
   NOTE: For the purposes of ONDRI, an in-house designed Python script is used to perform ACMG classification on a semi-automated basis. Although not used for this study, InterVar[40] is a similarly designed tool that can be utilized in an analogous manner.
6. Sanger sequence any variants with a sequencing coverage of <30x and/or variants that have been identified in > 10% of the study cohort to validate that they are not sequencing artifacts[41].

## Representative Results

The methodologies described herein were applied to 528 participant DNA samples from individuals that have been enrolled in ONDRI. Samples were run on the ONDRISeq panel in 22 runs of 24 samples per run. Overall, sequencing data were determined to be of high quality with a mean sample coverage of 78 ± 13x and all individual runs expressed a mean sample coverage >30x. Further, on average, 94% of all target regions were covered at least 20x (**Table 1**).

A mean 95.6% of reads were mapped to the reference sequence and all ONDRISeq runs had >90% of reads mapped (**Table 1**). Of the mapped reads, 92.0% had a Phred score ≥Q30, with only one run having <80% of mapped reads meeting this quality metric. However, this run still displayed a mean coverage of 79x and 93% of target regions were covered at least 20x.

| Parameter | Mean (±sd) | Best performance | Poorest performance |
|---|---|---|---|
| Cluster Density ($x10^3/mm^2$) | 1424 (±269) | 1347 | 1835 |
| Total Reads ($10^6$) | 43.1 (±6.0) | 48.7 | 47.4 |
| Mapped Reads ($10^6$) | 40.1 (±6.0) | 47.1 | 25.7 |
| Mapped Reads (%) | 95.6 (±1.3) | 96.8 | 92.6 |
| Phred Quality Score ≥Q30 (%) | 92.0 (±6.0) | 92 | 68.3 |
| Sample Coverage (x) | 78 (±13) | 99 | 51 |

**Table 1: Sequencing quality metrics for 22 runs on ONDRISeq.**

**Case Study: Identification of rare variants in a PD patient.**

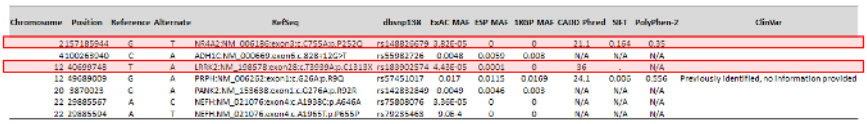
To demonstrate the utility of our targeted NGS workflow, we present the example of a 68 year-old, male, Parkinson's disease patient. The DNA sample was run on the NGS desktop instrument (**Table of Materials**) using the ONDRISeq panel alongside 23 other ONDRI samples. The run displayed a cluster density of 1,555 x $10^3/mm^2$. The patient's particular sample displayed a mean coverage of 76x, with 93.9% of the target regions covered at least 20x.

After performing variant calling and annotation with the custom bioinformatics workflow, the patient was found to harbor 1351 variants within the exons and surrounding 250 bp of the 80 genes included on the ONDRISeq panel. However, the ANNOVAR pipeline was able to reduce the number of variants by considering variant sequence ontology and MAF, as described above. This produced a list of seven variants that underwent manual curation (**Figure 3**). From these seven variants, two were identified as having possible clinical significance. This process is specific to the needs of ONDRI and was done by identifying those that are relatively rare in the general population and are nonsynonymous in ontology thereby causing a change in the protein. Whether the variant had been previously associated with disease, the *in silico* predictions of deleteriousness to the protein and the ACMG pathogenicity classification of the variants were also utilized in this process.

The first identified from the reduced list was a heterozygous variant, namely *LRRK2*:c.T3939A, resulting in the nonsense variant p.C1313*. *LRRK2* encodes the protein Leucine-Rich Repeat Kinase 2, which possesses both GTPase and kinase activity[42]. Further, mutations within this gene are known to be among the leading causes of familial Parkinson's disease[43]. This variant introduces a premature stop codon within *LRRK2*, thereby losing amino-acid residues 1,314–2,527. This prevents the translation of the protein's Ras of complex proteins (Roc), C-terminal of Roc (COR), and protein kinase domains, which are involved in functioning as an atypical Rho GTPase, GTP binding protein, and protein kinase, respectively, and was predicted to be damaging by the *in silico* analysis generated by CADD (CADD Phred = 36). This variant is also rare with a MAF of 0.004% and 0.01% in ExAC and ESP, respectively, and is absent from the 1000G database. Additionally, this is the only patient out of all 528 sequenced who carries this variant, which is novel since it has not been previously described in disease mutation databases (**Table of Materials**). The confidence of the variant call was confirmed by its deep coverage of 109x. Finally, the variant was assessed with the AMCG Standards and Guidelines for pathogenicity and was classified as being pathogenic.

The patient also carried a second heterozygous variant, *NR4A2*:c.C755A, resulting in the missense change p.P252Q. The protein encoded by *NR4A2*, Nuclear Receptor Subfamily 4 Group A Member 2, is a transcription factor involved in the generation of dopaminergic neurons[44] and mutations within this gene have been previously associated with Parkinson's disease[45]. The substitution of the non-polar proline to the polar glutamine was predicted to be damaging by the *in silico* prediction analysis generated by CADD (CADD Phred = 21.1), but not by the analysis generated by SIFT or PolyPhen-2. The variant is rare, with a MAF of 0.004% in ExAC and absence from both ESP and 1000G. The variant was also identified in an ONDRI participant diagnosed with vascular cognitive impairment, but has not been previously described in disease mutation databases. This variant had coverage of only 18x, however, Sanger sequencing will be performed in order to ensure its validity within the sequence. Finally, the variant was determined to be of uncertain significance when assessed with the ACMG Standards and Guidelines for pathogenicity.

The ONDRISeq panel and bioinformatics pipeline is also able to determine the *APOE* genotype of each sample. This patient was determined to have the *APOE* genotype E3/E3.



| Chromosome | Position | Reference | Alternate | RefSeq | dbsnp138 | ExAC MAF | ESP MAF | 1KGP MAF | CADD Phred | SIFT | PolyPhen-2 | ClinVar |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 157185944 | G | T | NR4A2:NM_006186:exon3:c.C755A:p.P252Q | rs148836679 | 3.83E-05 | 0 | 0 | 21.1 | 0.164 | 0.35 | |
| 4 | 100269040 | C | A | ADH1C:NM_000669:exon5:c.828-12G>T | rs59982726 | 0.0048 | 0.0036 | 0.008 | N/A | N/A | N/A | |
| 12 | 40689748 | T | A | LRRK2:NM_198578:exon28:c.T3939A:p.C1313X | rs189901574 | 4.42E-05 | 0.0001 | 0 | 36 | | N/A | |
| 13 | 49089009 | G | A | PRPH:NM_006262:exon1:c.G268A:p.R90Q | rs57451017 | 0.017 | 0.0115 | 0.0169 | 24.1 | 0.003 | 0.556 | Previously identified, no information provided |
| 20 | 3870023 | C | A | PANK2:NM_153638:exon1:c.C276A:p.R92R | rs142883840 | 0.0049 | 0.0046 | 0.003 | N/A | N/A | N/A | |
| 22 | 19885567 | A | C | NEFH:NM_021076:exon4:c.A1938C:p.A646A | rs75808076 | 3.30E-05 | 0 | 0 | N/A | N/A | N/A | |
| 22 | 29885504 | A | T | NEFH:NM_021076:exon4:c.A1955T:p.P652P | rs79255468 | 9.0E-4 | 0 | 0 | N/A | N/A | N/A | |

**Figure 3: Example of a reduced output from ANNOVAR displaying manually curated, annotated variants.** The reduced ANNOVAR output from the case study of a 68 year old, male, patient with Parkinson's disease. Annotated variants are curated to identify those that are most likely to be of clinical significance, as denoted by the red boxes. Please click here to view a larger version of this figure.

## Discussion

In the path from DNA sample extraction to identifying variants that may be of interest when considering a patient's diagnosis, disease progression, and possible treatment options, it is important to recognize the multifarious nature of the methodology required for both sequencing and proper data processing. The protocol described herein is an example of the utilization of targeted NGS and subsequent bioinformatic analysis essential to identify rare variants of potential clinical significance. Specifically, we present the approach taken by the ONDRI genomics subgroup when using the ONDRISeq custom-designed NGS panel.

It is recognized that these methods were developed based on a specific NGS platform and that there are other sequencing platforms and target enrichment kits that may be used. However, the NGS platform and desktop instrument (**Table of Materials**) was chosen based on its early US Food and Drug Administration (FDA) approval[46]. This authorization reflects the high-quality sequencing that can be performed with the NGS protocols of choice and the reliability that can be placed on the sequencing reads.

Although obtaining accurate sequencing reads with the depth of coverage is very important, the bioinformatics processing required for final rare variant analysis is vital and can be computationally intensive. Due to the many sources of errors that may occur within the sequencing process, a robust bioinformatics pipeline must correct for the various inaccuracies that can be introduced. They may arise from misalignments in the mapping process, amplification bias introduced by PCR amplification in the library preparation, and the technology producing sequencing artifacts[47]. No matter the software used to perform read mapping and variant calling, there are common ways to reduce these errors including local realignment, removal of duplicate mapped reads, and setting proper parameters for quality control when calling variants. Additionally, the parameters chosen during variant calling may vary based on what is most appropriate for the study at hand[11]. The minimum coverage and quality score of a variant and the surrounding nucleotides that were applied herein were chosen as to create a balance between appropriate specificity and sensitivity. These parameters have been validated for the ONDRISeq panel based on variant calling concordance with three separate genetic techniques, as previously described, including: 1) chip-based genotyping; 2) allelic discrimination assay; and 3) Sanger sequencing[9].

Following accurate variant calling, in order to determine those of potential clinical significance, annotation and curation are essential. Due to its open access platform, ANNOVAR is an excellent tool for both annotation and preliminary variant screening or elimination. Beyond being easily accessible, ANNOVAR can be applied to any VCF file, no matter what sequencing platform is used, and is customizable based on the needs of the research[26].

After annotation, variants must be interpreted to determine if they should be considered to be of clinical significance. Not only does this process become complex, but it is often prone to subjectivity and human error. For this reason, the ACMG has set guidelines to assess the evidence for pathogenicity of any variant. We apply a non-synonymous, rare variant-based manual curation approach, which is constructed based on these guidelines and safeguarded by individually assessing each variant that is able to pass through the pipeline with a custom-designed Python script that classifies the variants based on the guidelines. In this way, each variant is assigned a ranking of pathogenic, likely pathogenic, uncertain significance, likely benign, or benign, and we are able to add standardization and transparency to the variant curation process. It is important to

recognize that the specifics of variant curation, beyond the bioinformatics pipeline, will be individualized based on the needs of the research, and was therefore beyond the scope of the methodologies presented.

Although the methods presented here are specific to ONDRI, the steps described can be translated when considering a large number of constitutional diseases of interest. As the number of gene associations increase for many phenotypes, targeted NGS allows for a hypothesis driven approach that can capitalize on the previous research that has been done in the field. Yet, there are limitations to targeted NGS and the methodology presented. By only focusing on specific regions of the genome, the areas of discovery are limited to novel alleles of interest. Therefore, novel genes or other genomic loci beyond those covered by the sequencing targets, which could be revealed with WGS or WES approaches, will not be identified. There are also regions within the genome that can be difficult to accurately sequence with NGS approaches, including those with a high degree of repeated sequences[48] or those that are rich in GC content[49]. Fortunately, when utilizing targeted NGS, there is a priori a high degree of familiarity with the genomic regions being sequenced, and whether these might pose technical challenges. Finally, detection of copy number variants from NGS data at present is not standardized[50]. However, bioinformatics solutions to these concerns may be on the horizon; new computational tools may help to analyze these additional forms of variation in ONDRI patients.

Despite its limitations, targeted NGS is able to obtain high-quality data, within a hypothesis-driven approach, while remaining less expensive than its WGS and WES counterparts. Not only is this methodology appropriate for efficient and directed research, the clinical implementation of targeted NGS is growing exponentially. This technology is being used to answer many different questions regarding the molecular pathways of various diseases. It is also being developed into an accurate diagnostic tool at relatively low cost when opposed to WES and WGS. Even when compared to the gold-standard Sanger sequencing, targeted NGS can outcompete in its time- and cost-efficiency. For these reasons, it is important for a scientist or clinician who receives and uses NGS data, for instance, delivered as text in a laboratory or clinical report, to understand the complex "black box" that underlies the results. The methods presented herein should help users understand the process underlying the generation and interpretation of NGS data.

## Disclosures

The authors have nothing to disclose.

## Acknowledgements

We would like to thank all ONDRI participants for their consent and cooperation with our study. Thank you to the ONDRI investigators (www.ONDRI.ca/people), including our lead investigator (MJS), and the ONDRI governing committees: the executive committee, steering committee, publication committee, recruiting committee, assessment platforms, and project management team. We also thank the London Regional Genomics Centre for their technical expertise. AAD is supported by the Alzheimer Society of London and Middlesex Masters Graduate Research Scholarship. SMKF is supported by the ALS Canada Tim E. Noël Postdoctoral Fellowship.

## References

1. Metzker, M. L. Sequencing technologies - the next generation. *Nat Rev Genet.* **11** (1), 31-46 (2010).
2. Mardis, E. R. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet.* **9** 387-402 (2008).
3. Shendure, J., & Ji, H. Next-generation DNA sequencing. *Nat Biotechnol.* **26** (10), 1135-1145 (2008).
4. Sanger, F., Nicklen, S., & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A.* **74** (12), 5463-5467 (1977).
5. Farhan, S. M. K., & Hegele, R. A. Exome Sequencing: New Insights into Lipoprotein Disorders. *Current Cardiology Reports.* **16** (7) (2014).
6. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A.* **106** (45), 19096-19101 (2009).
7. Mardis, E. R. DNA sequencing technologies: 2006-2016. *Nat Protoc.* **12** (2), 213-218 (2017).
8. Farhan, S. M. *et al.* The Ontario Neurodegenerative Disease Research Initiative (ONDRI). *Can J Neurol Sci.* **44** (2), 196-202 (2017).
9. Farhan, S. M. K. *et al.* The ONDRISeq panel: custom-designed next-generation sequencing of genes related to neurodegeneration. *NPJ Genom Med.* (16032), 1-11 (2016).
10. El-Metwally, S., Hamza, T., Zakaria, M., & Helmy, M. Next-generation sequence assembly: four stages of data processing and computational challenges. *PLoS Comput Biol.* **9** (12), e1003345 (2013).
11. Yohe, S., & Thyagarajan, B. Review of Clinical Next-Generation Sequencing. *Arch Pathol Lab Med.* (2017).
12. *Gentra Puregene Handbook.* 4th edn. Qiagen (2014).
13. *Spectrophotometer V3.5 User's Manual.* NanoDrop Technologies, Inc. (2007).
14. *Qubit 2.0 Fluorometer User Manual.* Vol. Q32866. Invitrogen by Life Technologies. (2010).
15. *Nextera Rapid Capture Enrichment Guide.* Vol. 15037436 v01. Illumina, Inc. (2016).
16. *Nextera Rapid Capture Enrichment Reference Guide.* Vol. 15037436 v01. Illumina, Inc. (2016).
17. *MiSeq Reagent Kit v3 Reagent Preparation Guide.* Vol. 15044932 Rev. B. Illumina, Inc. (2013).
18. *MiSeq System Guide.* Vol. 15027617 v01. Illumina, Inc. (2015).
19. *BaseSpace Sequence Hub.* <https://basespace.illumina.com/dashboard> (2017).
20. *Agilent High Sensitivity DNA Kit Guide.* Vol. G2938-90321 Rev. B. Agilent Technologies (2013).
21. *MiSeq System Denature and Dilute Libraries Guide.* Vol. 15039740 v01. Illumina, Inc., (2016).
22. *System Specification Sheet: MiSeq System.* Illumina, Inc. (2016).
23. *BaseSpace Sequence Hub Help Center.* <https://help.basespace.illumina.com/> (2017).
24. *Genomics Workbench 10.1.1 User Manual.* Qiagen (2017).
25. Ebbert, M. T. *et al.* Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics.* **17 Suppl 7** 239 (2016).

26. Wang, K., Li, M., & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38** (16), e164 (2010).
27. Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44** (D1), D733-745 (2016).
28. Kitts, A., Phan, L., Ward, M., & Bradley Holmes, J. *The Database of Short Genetic Variation (dbSNP)*. (ed Bethesda (MD)) National Center for Biotechnology Information (US), (2013).
29. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature.* **536** (7616), 285-291 (2016).
30. *Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP).* <http://evs.gs.washington.edu/EVS/> (2017).
31. Auton, A. *et al.* A global reference for human genetic variation. *Nature.* **526** (7571), 68-74 (2015).
32. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44** (D1), D862-868 (2016).
33. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet.* **46** (3), 310-315 (2014).
34. Kumar, P., Henikoff, S., & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc.* **4** (7), 1073-1081 (2009).
35. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods.* **7** (4), 248-249 (2010).
36. Bertram, L., McQueen, M. B., Mullin, K., Blacker, D., & Tanzi, R. E. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat Genet.* **39** (1), 17-23 (2007).
37. Xiong, H. Y. *et al.* The human splicing code reveals new insights into the genetic determinants of disease. *Science.* **347** (6218) (2015).
38. Desmet, F. O. *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* **37** (9), e67 (2009).
39. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med.* **17** (5), 405-424 (2015).
40. Li, Q., & Wang, K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *Am J Hum Genet.* **100** (2), 267-280 (2017).
41. Yang, Z. L., & Sun, G. L. High-frequency, low-coverage "false positives" mutations may be true in GS Junior sequencing studies. *Scientific Reports.* **7** (2017).
42. Gandhi, P. N., Wang, X., Zhu, X., Chen, S. G., & Wilson-Delfosse, A. L. The Roc domain of leucine-rich repeat kinase 2 is sufficient for interaction with microtubules. *J Neurosci Res.* **86** (8), 1711-1720 (2008).
43. Goldwurm, S. *et al.* The G6055A (G2019S) mutation in LRRK2 is frequent in both early and late onset Parkinson's disease and originates from a common ancestor. *J Med Genet.* **42** (11), e65 (2005).
44. Caiazzo, M. *et al.* Direct generation of functional dopaminergic neurons from mouse and human fibroblasts. *Nature.* **476** (7359), 224-227 (2011).
45. Grimes, D. A. *et al.* Translated mutation in the Nurr1 gene as a cause for Parkinson's disease. *Mov Disord.* **21** (7), 906-909 (2006).
46. Collins, F. S., & Hamburg, M. A. First FDA authorization for next-generation sequencer. *N Engl J Med.* **369** (25), 2369-2371 (2013).
47. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics.* **43** 11 10 11-33 (2013).
48. Treangen, T. J., & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet.* **13** (1), 36-46 (2011).
49. Shin, S., & Park, J. Characterization of sequence-specific errors in various next-generation sequencing systems. *Mol Biosyst.* **12** (3), 914-922 (2016).
50. Povysil, G. *et al.* panelcn.MOPS: Copy-number detection in targeted NGS panel data for clinical diagnostics. *Hum Mutat.* **38** (7), 889-897 (2017).